SYSTEM IMPROVEMENT AND REPORTING DIVISION

Grade Level of Achievement Reporting Initiative

From Conceptualization to Pilot to Implementation

Fall 2005

Alberta
EDUCATION

For further information, contact

System Improvement and Reporting Division
9th Floor, Commerce Place
10155 102 Street
Edmonton, Alberta T5J 4L5

Telephone (780) 422-8671
Toll free in Alberta by dialing 310-0000
Fax: (780) 422-8345
Email: sig@edc.gov.ab.ca

Table of Contents

Introduction

The underlying philosophy in the Grade Level of Achievement Reporting initiative is that it is possible to balance the decision-making or management information needs of administrators and teachers while simultaneously supporting the learning needs of students and the information needs of government. To this end, following an extensive study of alternative approaches to program evaluation that would go well "beyond MIRS", a pilot project was undertaken in the 2002-03 school year wherein Grade Level of Achievement (GLA) data was collected from 202 schools for 51,816 students, to assess the validity, reliability and ultimately the utility of collecting GLA data for evaluating program impacts.

An objective of this paper is to expand awareness and understanding of the Grade Level of Achievement Reporting initiative being phased in over a three year period beginning in the 2005-06 school year. The paper will provide a brief synopsis of the theoretical underpinnings of the Grade Level of Achievement Reporting initiative, as well as a description of the pilot project or case study. It will then show that the data collected therein has sufficient reliability and validity so as to make GLA data based on numerous and more dynamic observations over time aggregated up from the classroom level, a viable option for judging program impacts at a provincial level, and will consider some future implications stemming from the usefulness of the data.

BACKGROUND

Alberta Learning began collecting data based on Management Information Reporting Schedules (MIRS) information in 1998/99 as a means of monitoring the effectiveness of specific funded programs. The schedules requested information from schools aggregated by jurisdictions on the program inputs and/or the performance of students who were receiving additional funding for programs such as special education, early literacy, and English as a Second Language to name a few. An objective of the MIR schedules was to inform Alberta Education and jurisdiction decision-making by analyzing achievement gains for subgroups of the student population being served by specific programs, and to provide accountability related to the funding structure.

Despite the fact that this was an important attempt to help insure students in these funding categories were not falling between the cracks, or falling behind, the utility of the data generated was marginal in terms of its applicability to re-designing program delivery owing to a lack of controls that would have helped to ensure data quality and to demonstrate causality. In addition to these limitations, was the burden on schools of reporting these ad hoc, divergent data requests on a yearly basis. The MIR schedules were seen as another new activity administrators had to complete in an already busy job. Reporting MIRS did not "piggy-back" well with the other reporting activities schools direct as part of their core business, and the onerous nature of the reporting coupled with limited utility of the data for the intended purposes led in part to the cessation of MIRS reporting in 2002. While this alleviated the reporting burdens, the

management information needs for program evaluation remained at the provincial, jurisdiction and school levels[1].

From the provincial perspective, the need to monitor student progress pertaining to the specially funded programs has changed somewhat because of changes in the funding formula. However, the need to evaluate programs and their effectiveness for all students and particular subsets of students, at a provincial, jurisdictional and school level remains an ongoing concern, especially in light of provincial priorities such as improving the high school completion rate. The Grade Level of Achievement Reporting initiative grew out of these needs, but it was also seen as an opportunity to develop teacher capacity to do good classroom assessment work, to improve pedagogy by linking assessment with instructional decision-making, and as an approach to better engage teachers, administrators, students and parents in formative assessments in ways that complement summative assessments.

Theoretical Underpinnings

Educational accountability has become firmly entrenched in learning systems across North America and Europe. Policy-makers and educators alike can appreciate the benefit of organizing student improvement around the concept of making education goals known and using a broad range of measures of the progress towards those goals in the hopes that doing so will improve student achievement[2]. Effective accountability is premised upon creating open education systems that state outright the goals educators wish to accomplish, supplemented by the right to ask what went wrong if the goals are not achieved. For the most part, this is a logical process where setting standards, ensuring some commonality in reporting by employing standardized assessments and reserving the right to hold people accountable for their actions if the results are not as expected, is not an unreasonable set of parameters for a publicly funded education system. However, is this external need for information on student performance compatible with internal school-based improvement approaches?

Many government-level models for improving student achievement are based on non-education sectors and their teachings have been expanded to the education world. Education like any other sector can benefit from accountability in this form, but the challenge lies in providing opportunities for teachers to see benefit and acquire ownership for the accountability processes including data and the sense making opportunities associated with that data (Louis, Febey and Schroeder, 2005). The disconnect occurs when the primary participants in classroom assessment, the teachers, see a gap between their assessment efforts and the external assessment initiatives of governments. More specifically, the amount of accountability effort in education measured using formative, well-rounded classroom assessment methods (assessment for learning) need to be balanced with summative standardized assessments (assessment of learning) so that comprehensive and compatible information is available to more fully inform decisions around what is working for students (Stiggins, 2001) (Earl and Katz, 2002).

---

[1] See Alberta Learning, *MIRS NEXT GENERATION: Design Principles for a Learner Results Database for Improved Program Evaluation*, August, 2001.

[2] See for example the National Centre for Educational Accountability in the United States. Their mission states firmly that they aim to promote student achievement by improving state data collection to improve decision-making. http://www.nc4ea.org/index.cfm?pg=about_us

Educators such as Bloom (1980), Stiggins (2001) and Reeves (2004) are quick to point out that students do not improve most when only assessment of learning or summative assessment techniques are employed, but when assessment for learning or formative assessment techniques are used in an appropriate balance with the summative forms. Yet, media around the world have a tendency to place an exclusive emphasis on summative assessments in their coverage of accountability.

Standardized tests and other forms of summative assessment undoubtedly have a place in education accountability owing to the high quality of these instruments and the valid and reliable nature of the data they provide. However, standardized testing is only one piece of the accountability puzzle[3].

Hence, the Grade Level of Achievement Reporting initiative is partly premised on the notion that summative, assessment of learning models should not be viewed as the only data sources available to policy makers. However, the collection of GLA data is not intended for use in Alberta Education's Accountability Pillar. It is intended to be used in monitoring program impacts at the provincial level, but could add considerable depth and meaning to data compiled at the school and jurisdiction levels especially when considered in relationship to appropriate data from provincial achievement tests.

The other value directing the Grade Level of Achievement Reporting initiative is the view that engaging educators at the classroom level in ongoing assessment for learning as a basis for informing GLA reporting is absolutely vital to sound pedagogy and teacher professionalism. Accordingly, there is expectedly a willingness among education professionals to embrace the opportunity to relate classroom GLA data to broad-based assessment methods recognizing that, as Reeves (2004) states:

> ...the judgment of the classroom teacher is an integral part of constructive accountability....Only when accountability, standards, and assessment are fully integrated at the classroom level will we achieve the potential for fairness, equity of opportunity, and improved academic achievement that *teaching professionals crave* and society demands.

An integrated student achievement database would flesh out and balance the data generated by the provincial achievement testing program with teacher-based assessment of student achievement and provide a much more dynamic, complete and enriched picture of student curricular-based learning while enhancing the professional role of teachers in this process. Furthermore, the creation and maintenance of such a database would not represent a big leap over existing work that teachers do, but by creating a system to routinely collect and aggregate student grade level of achievement data some significant gaps in our knowledge about what is working for students could be illuminated.

Given the absence of a reporting mechanism for key information informing program evaluation decisions that was created when the MIR schedules were discontinued, an opportunity was

[3] See Burger and Krueger (2003) "A Balanced Approach to High-Stakes Achievement Testing: An analysis of the literature with policy implications".

recognized to attempt a data collection initiative based on classroom generated data. Doing so, it was hoped, would demonstrate the following:

1) Grade level of achievement data driven by formative, assessment for learning methods is a reasonable source of information, with acceptable concurrent and predictive validity, that is useful for informing summative judgments of program impacts.

2) Classroom generated grade level of achievement data adds value and contributes to the Ministry's and the public's knowledge and understanding of the existing data collected for schools and jurisdictions.

3) The process of generating grade level of achievement data for reporting has a positive impact on teacher professional growth and pedagogy.

The remainder of this paper will attempt to address the above by first describing the results of a quantitative examination of the pilot GLA data in the next section, before focusing on the qualitative implications of the pilot project.

Pilot Design and Description of Data

Two hundred and two schools from 4 jurisdictions submitted grade level of achievement data at the end of the 2002-03 school year for 51,816 students[4]. The fields collected included student name (surname and given name), Alberta student number, and enrolled grade. Enrolled Grade was defined as the grade to which the student was assigned. Typically there is a strong relationship between a student's age, peer group and enrolled grade.

GLA was collected for all students on graded curriculum, including those with special needs, in the following fields where applicable:

- GLA in English Language Arts
- If applicable - GLA in French Language Arts
- GLA in Mathematics

Grade Level of Achievement was defined as the grade level expressed as a whole number in relationship to the learning outcomes defined in the Program of Studies that teachers judged the student to have achieved at the end of the 2003/04 school year. Some school boards apply a standard test or battery of tests to help determine the grade level of achievement. If that was true for the school submitting the data, teachers were asked to consider that assessment in relationship to the full range of assessment information available to them, including classroom assessment marks, in making a professional judgment of the student's grade level of achievement.

For students with special needs who were not on a graded curriculum (i.e. not based on the Programs of Study), teachers were asked to check one of the following descriptions that best described the goals in the student's Individualized Program Plan (IPP) that had been met. If

[4] The majority of the data (98.2%) was submitted by the Edmonton Public School District. This number represents the total number of student records and includes those where GLA data was missing for either math or language arts.

goals were met, teachers were asked to respond, "YES". If the goals were not met teachers were instructed to respond "NO", and if not applicable they were instructed to respond "N/A" (Note: this reporting structure has since been changed to include an expanded set of descriptors and the wider range of responses).

- *Student has met IPP goals and objectives that address communication skills.*
- *Student has met IPP goals and objectives that address functional skills.*

"Not on Graded Curriculum" was meant to indicate that the student's program was restricted to learning outcomes that were significantly different from the provincial curriculum defined in the Program of Studies and were specifically selected to meet the student's special needs as defined in the *Standards for Special Education* (Alberta Learning, 2002).

"Communication Skills" referred to the development of expressive and or receptive communication. This could be verbal communication and/or alternative modes of communication. "Functional Skills" referred to skills that would assist the student in developing independence in the home, school and community.

The following illustrative examples were provided on the GLA data collection form to help increase the reliability of the submitted data:

- Student A is enrolled in grade 4. Her Language Arts program is based on the grade 4 learning outcomes defined in the English Language Arts k-9 Program of Studies. The full range of assessment results for Student A demonstrates she has achieved the outcomes for grade 4 so the data is entered, "achieved grade 4."
- Student B is enrolled in grade 8. He has been coded as having a mild learning disability. His Math program is based on the grade 6 learning outcomes defined in the Math k-9 Program of Studies. The full range of assessment results for Student B demonstrates he has achieved the outcomes for grade 6 so the data is entered, "achieved grade 6."
- Student C is enrolled in grade 2. He has been coded as having a severe learning disability. His Language Arts program is based on developing language arts readiness skills and on some of the grade 1 learning outcomes defined in the English Language Arts k-9 Program of Studies. The full range of assessment results for Student C demonstrates he has not achieved all of the learning outcomes for grade 1 so the data is entered, "not yet 1."
- Student D is enrolled in grade 3. She has been coded as having multiple severe disabilities and works with a full time aide. Her program is based completely on learning objectives that are below the grade 1 learning outcomes defined in the Math or English Language Arts k-9 Program of Studies. Her Individualized Program Plan defines communication and functional skill outcomes designed to develop independent living skills. All of the IPP outcomes for the current school year have been achieved so the data is entered in Part C, "Yes" for both communication and functional skills.

The Alberta student number was used by Alberta Education staff to append data fields such as Provincial Achievement Test (PAT) results (both raw scores and achievement levels), student age, number of school registrations, any additional student codes associated with the student, and school starting date. Individual student identifiers were replaced with a discreet Grade Level of Achievement Reporting ID, leaving no personal identifiers in the dataset.

Limitations of the Data

When analyzing the data, the following limitations were noted.

- Nearly 98% of the data submitted was from one jurisdiction, which has been collecting GLA data for several years.
- Of the total 51,816 records, 1,456 (approximately 2.8%) had no GLA data submitted for English Language Arts, and 1,358 (approximately 2.6%) had no GLA data submitted for Math.
- Of the 934 records submitted by other jurisdictions, 69 of the records submitted had no English Language Arts GLA data. However, 57 of these had IPP data submitted, meaning there was only 1.5% of the valid population with no English Language Arts GLA data. 62.3% of the same population had no data submitted for Math GLA.
- IPP data were submitted for only 57 students, meaning there were only 57 students not on a graded curriculum.

The data are approximately equally distributed by enrolled grade with 10-11% of the overall students from grade 1 to grade 9 in each grade cohort. If the students were distributed exactly evenly in each grade, we would expect 5671 students per grade, or approximately 11%. The table below shows the distribution by enrolled grade.

Enrolled Grade Distribution

Enrolled Grade	Frequency	Valid Percent
1	5228	10.2%
2	5385	10.6%
3	5559	10.9%
4	5661	11.1%
5	5711	11.2%
6	5831	11.4%
7	6272	12.3%
8	6099	11.9%
9	5292	10.4%
Sub-Total	51038	100.0%
10	778	
Total	51816	

An irregularity was apparent in that there were 778 students in the database with an enrolled grade of 10, but a GLA of 9. As the data was collected only for students in grades 1 to 9, these 778 were treated as anomalies and not used in any analyses by enrolled grade. They were however used as valid cases in analyses that were not grade specific.

88.9% of the students were Non-Coded, meaning they had not been identified as having any type of special need. Approximately 2.1% (1,080) of the students had severe codes (codes 40 through 49), 7.8% (4066) had mild/moderate codes, and 1.2% were coded as gifted/talented (609).

Recoded Expanded Code Variable into Groups- Population Parameters	Frequency	Percent
Non-Coded	46061	88.9%
Severe Disabilities (Code 40 thru 49)	1080	2.1%
Mild/Moderate Disabilities	4066	7.8%
Gifted and Talented	609	1.2%
Total	51816	100%

Correlations between GLA and Enrolled Grade by Sub-Groups of the Population-

Correlations between the students' GLAs and enrolled grades were calculated using Spearman's rho to determine the "goodness of fit" between GLA and enrolled grade. The correlation between the two variables reflects the degree to which the variables are related, or the degree to which they "move" together. A high positive correlation coefficient results when an increase in one variable is mirrored by the same increase in another. Spearman's is used specifically to measure ordinal level data in that it first converts the data to rank orders before correlating. As was expected, GLA was highly correlated to enrolled grade, meaning the enrolled grade of a student typically matches their GLA. Again as predicted, this relationship was strongest for the sub-population that had no group codes attached to their records, or non-coded students, while students with severe disabilities had the lowest correlation between GLA and enrolled grade while students with mild/moderate codes were between these two groups. This relationship was true when testing both Math and English Language Arts GLA against enrolled grade.

English GLA

Recoded Expanded codes into groups	*Correlation Coefficient*
Non-Coded	.992(**)
Severe Disabilities (Code 40 thru 49)	.788(**)
Mild/Moderate Disabilities	.852(**)
Gifted and Talented	.999(**)

Math GLA

Recoded Expanded *codes into groups*	*Correlation Coefficient*
Non-Coded	.995(**)
Severe Disabilities (Code 40 thru 49)	.857(**)
Mild/Moderate Disabilities	.868(**)
Gifted and Talented	.974(**)

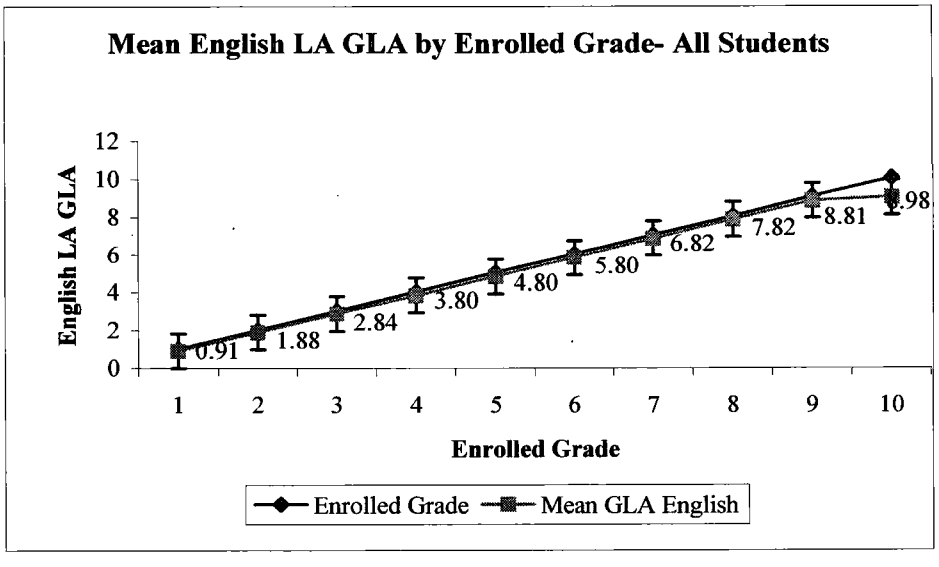
** Correlation is significant at the 0.01 level (2-tailed).

The following series of graphs show GLA by enrolled grade for all students as well as sub-populations of students, in English Language Arts.

The mean GLA is plotted against the enrolled grade to show the degree to which students' GLA reflect their enrolled grade. Additionally, a trend line was plotted for each graph using the formula y = bx + a, where y is the dependent variable, **b** is the slope, **x** is the independent variable and **a** is the y-intercept, or the value at which the line would cross the y-axis. B indicates the amount of increase in the dependent variable when the independent variable is increased by 1.

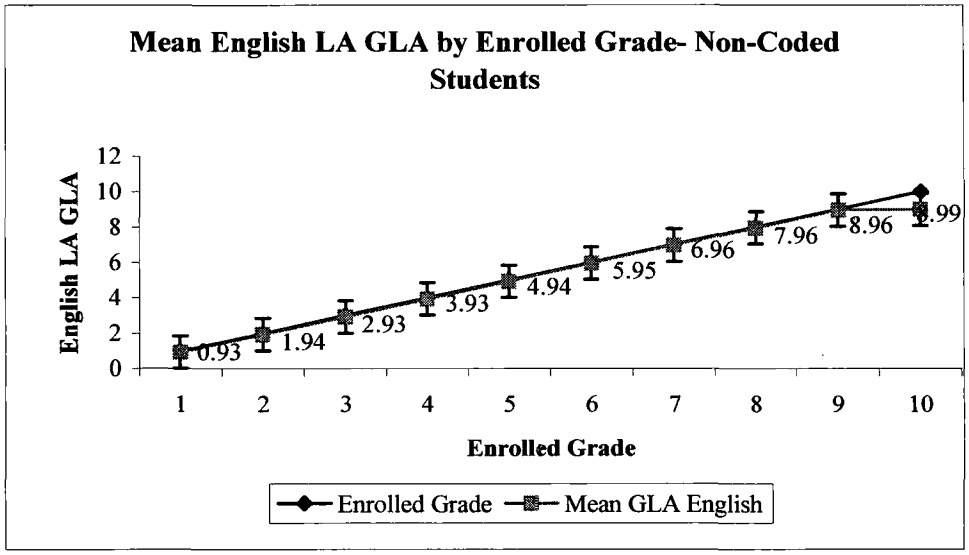
In the "Mean English LA GLA by Enrolled Grade- All Students" graph, the slope (b) is .9546, meaning we can expect GLA to increase by roughly .95 when the enrolled grade is increased by 1. In other words, this is a nearly perfect positive correlation. For complete frequency tables of GLA compared to enrolled grade see Appendix 2 of the Full Technical Report.

All Students- Entire Grade Level of Achievement Reporting Database



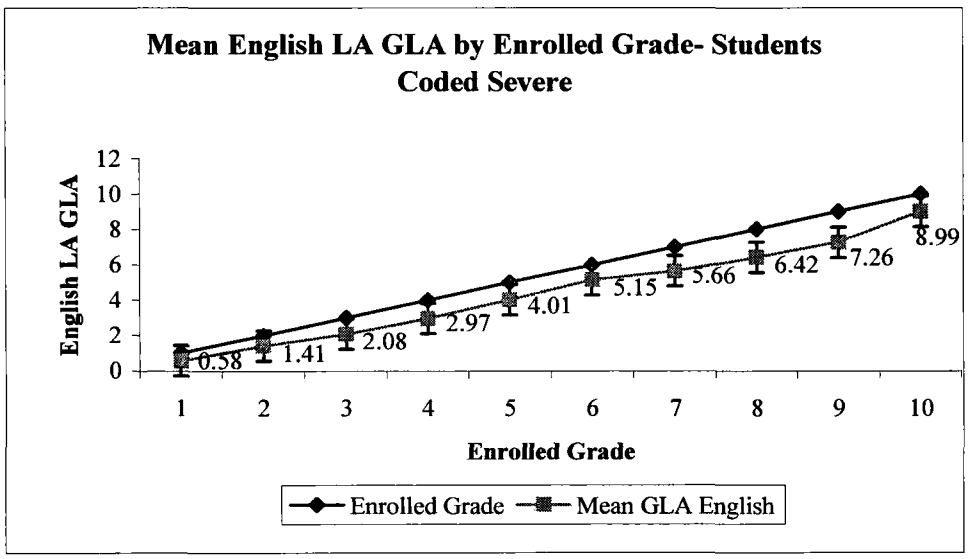
Above mean GLA line is y = 0.9546x + 0.0435
Formula for line is y = bx + a

Non-Coded Students



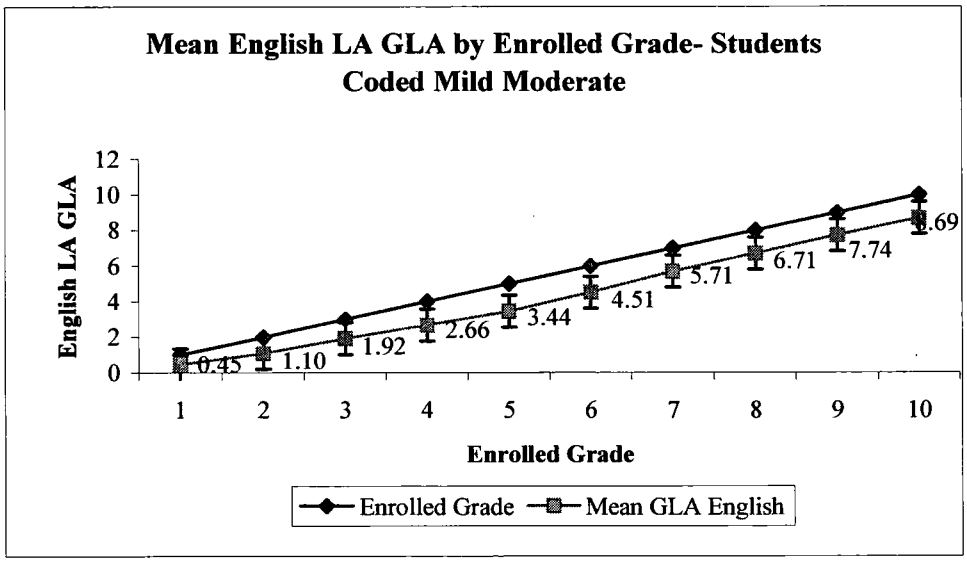
Above mean GLA line is y = .9512x + 0.1175

Students with Severe Codes



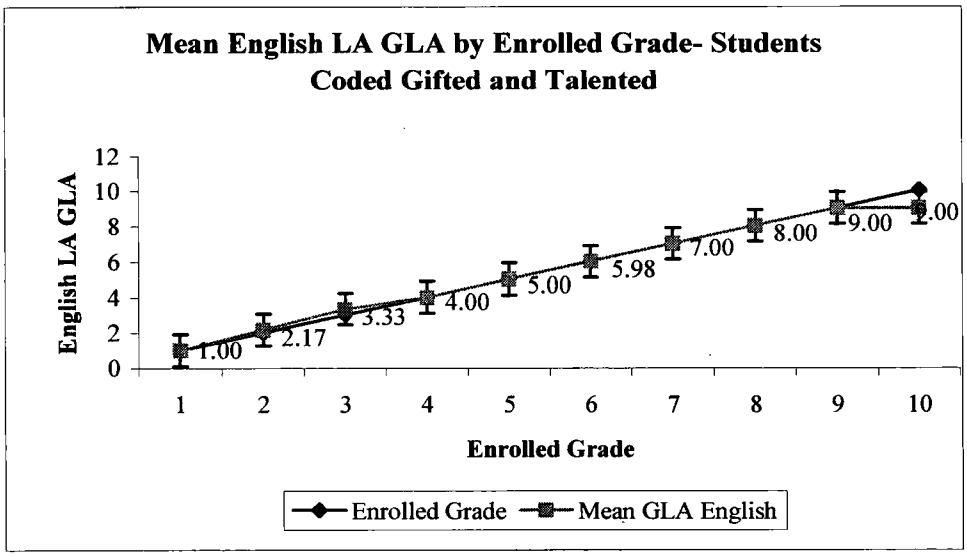
Above mean GLA line is y = .8946x - 0.465

Students with Mild Moderate Codes



Above mean GLA line is y = .9385x - 0.8692

Gifted and Talented Students



Above mean GLA line is y = .9281x + 0.3429

These graphs show that there is a good degree of face validity with the GLA data. For non-coded students the mean GLA in each grade matches the enrolled grade almost perfectly, and this is as expected. One would hypothesize that non-coded students' grade levels of achievement should match very precisely the grade they are enrolled in, and this is what the data show as the mean GLAs range from 0.07 to 0.02 decimals places below the enrolled grades in Math and English. Likewise, one would hypothesize that students with either mild moderate or severe

codes mean GLAs would not as precisely reflect the enrolled grade, and again this is what the data show. The mean GLAs in Math and English for students with severe codes range from 1.73 to .28 below enrolled grade, and mild moderate mean GLAs range from 1.56 to .46 below enrolled grade. Finally, for students coded gifted or talented, their mean GLA's in Math and English range from .02 below enrolled grade, to .33 above enrolled grade, which again reflects what one would reasonably expect given the tendency in many jurisdictions to enrich gifted students' programs as opposed to advancing a student ahead of their peer group.

GLA by PAT- *Comparisons using achievement levels*

In order to further examine the relationship between the Grade Level of Achievement Reporting data and provincial achievement tests (PATs), both PAT and GLA data were again re-coded into the dichotomous categories of either "Below Acceptable", or "At or Above Acceptable" for PATs; and "Below Grade Level" or "At or Above Grade Level" for GLA. These were then crosstabulated with the assumption being students who score at or above the acceptable level on PATs tend to be at or above grade level, and likewise those that score below acceptable tend to be below grade level, in the majority of cases.

The following tables resulted in supporting our hypothesis as 97%-99% of the students who are at grade level are also at or above the acceptable level.

		Grade Level of Achievement – English Language Arts		
		Below Grade Level	*At or Above Grade Level*	Total
PAT - Grade 3 English Language Arts	Below Accept.	33.5% (183)	66.5% (363)	546
	Accept. or Excellent	3.0% (133)	97% (4357)	4490
PAT - Grade 6 English Language Arts	Below Accept.	24.9% (162)	75.1% (488)	650
	Accept. or Excellent	2.3% (107)	97.7% (4482)	4589
PAT – Grade 9 English Language Arts	Below Accept.	15.6% (74)	84.4% (401)	475
	Accept. or Excellent	.9% (36)	99.1% (4070)	4106

		Grade Level of Achievement – Math		
		Below Grade Level	*At or Above Grade Level*	Total
PAT - Grade 3 Math	Below Accept.	19.3% (109)	80.7% (456)	565
	Accept. or Excellent	1.2% (51)	98.8% (4387)	4438
PAT - Grade 6 Math	Below Accept.	15.3% (92)	84.6% (508)	600
	Accept. or Excellent	.8% (37)	99.2% (4572)	4609
PAT – Grade 9 Math	Below Accept.	12.2% (112)	87.8% (803)	915
	Accept. or Excellent	.2% (7)	99.8% (3732)	3739

Gamma Analysis

All of the above observed relationships were significant when measured by Chi square[5]. Gamma values were subsequently calculated in order to determine the strength of the relationships.

Gamma is a proportional reduction in error (PRE) measure. In short, PREs measure the degree to which knowing the value of the independent variable will reduce error in predicting the value of the dependent variable. GLA was used as the independent measure, with PAT results being set as the dependent. In other words, Gamma provides us with a measure of the degree to which we will be able to predict a student's PAT achievement level, given their GLA.

The formula for Gamma is:

$$\gamma = \frac{Ns - Nd}{Ns + Nd}$$

Ns is the number of similar (concordant) pairs, and Nd is the number of dissimilar (discordant) pairs. To calculate Ns, each cell frequency is multiplied by the sum of the cell frequencies below and to the right of it, and then their products are summed. To calculate Nd, each cell frequency is multiplied by the sum of the cell frequencies above and to their right, and then their products are summed. For example, given the table below for Grade 9 English, Gamma would be calculated as follows:

[5] Chi square is a measure of the independence of two variables, and assesses the likelihood that any apparent relationship between the two is due to chance by comparing the observed frequencies to what would be expected if perfect independence existed.

Gamma- Grade 9 English

		Below Grade Level	*At or Above Grade Level*	Total
PAT - Grade 3 English Language Arts	Below Accept.	15.6% (74)	84.4% (401)	475
	Accept. or Excellent	.9% (36)	99.1% (4070)	4106

$$\gamma = \frac{Ns - Nd}{Ns + Nd}$$

$$Ns = 74 \times 4070$$
$$= 301{,}180$$

$$Nd = 36 \times 401$$
$$= 14{,}436$$

$$\gamma = \frac{301{,}180 - 14{,}436}{301{,}180 + 14{,}436}$$

$$\gamma = \frac{286{,}744}{315{,}616}$$

$$\gamma = .909$$

In layman's terms, this means knowing a student's Grade 9 English GLA level, gives us roughly a 91% chance of correctly predicting Grade 9 English LA PAT level. However, the above formula has a tendency to overstate the strength of a relationship when any cell has very low values, such as the acceptable PAT but below grade level cell in the grade 9 math data.

The following table lists the Gamma values for the relationships tested[6].

PAT by GLA- Grade and Subject	***Gamma***
Gr. 3 Eng., LA	.886
Gr. 6 Eng. LA	.866
Gr. 9 Eng. LA	.909
Gr. 3 Math	.907
Gr. 6 Math	.914
Gr. 9 Math	.973

[6] A similar analysis was conducted using just jurisdictions other than the main supplier of data. Owing to the smaller n's, it was only possible to calculate Gamma values for Grade 3 and Grade 6 English LA GLA by PAT. The resulting values were .950 and .687 respectively.

Analysis of Students Below Grade Level

In the Grade Level of Achievement Reporting pilot, it is possible to compare the ratings given by teachers through the GLA and by a standardized test through the provincial achievement test (PAT), in Grades 3, 6 and 9. In each case, it is possible to identify the students who are rated as below grade by their teachers (GLA) and those rated as below acceptable standard by the PAT.

One would expect some differences in the designation of individuals in the two ratings, since the teachers have an array of assessments available to do the rating whereas the PAT is a single pencil and paper test. However, since the objective of both methods is to measure how well a student is performing as compared to the learning outcomes in the Program of Studies, one would expect an overall positive relationship between the number of students identified as "below" by both methods.

An examination of the Grade Level of Achievement Reporting pilot data shows that this assumption departs most dramatically for grade 9 math within a general pattern where for both English Language Arts and Math fewer students are identified as "below" in the GLA ratings than are so identified in the PAT ratings. The following tables illustrate the differences:

Grade 3 ELA	Count	%
Wrote	5036	
Below on PAT	546	10.8%
Below on GLA	316	6.3%

Grade 3 Math	Count	%
Wrote	5003	
Below on PAT	565	11.3%
Below on GLA	160	3.2%

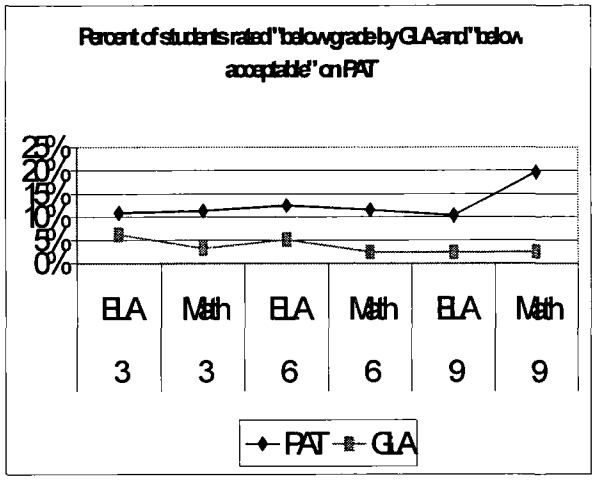
Grade 6 ELA	Count	%
Wrote	5239	
Below on PAT	650	12.4%
Below on GLA	269	5.1%

Grade 6 Math	Count	%
Wrote	5209	
Below on PAT	600	11.5%
Below on GLA	129	2.5%

Grade 9 ELA	Count	%
Wrote	4581	
Below on PAT	475	10.4%
Below on GLA	110	2.4%

Grade 9 Math	Count	%
Wrote	4652	
Below on PAT	915	19.7%
Below on GLA	119	2.6%

The above tables show a difference between the GLA and PAT ratings of "below", with the gap between the two ratings growing as grade levels increase. The increasing gap can also be shown graphically:



The above analysis seems contrary to the strong Gamma values, and as such, further study was undertaken.

Kendall's tau-b[7] values were calculated in the place of Gamma as a more conservative measure, using the formula:

$$\frac{Ns - Nd}{\sqrt{(Ns + Nd + Tx)(Ns + Nd + Ty)}}$$

where Ns and Nd are the same as Gamma, and Tx designates ties on the independent variable, and Ty designates ties on the dependent variable.

Again using the Grade 9 English values,
Ns = 74 x 4070
= 301,180

[7] Like Gamma, tau-b is also a PRE measure.

$$Nd = 36 \times 401$$
$$= 14{,}436$$

$$Tx = (74 \times 36) + (401 \times 4070)$$
$$Tx = 1{,}634{,}734$$

$$Ty = (74 \times 401) + (36 \times 4070)$$
$$Ty = 176{,}194$$

$$\text{Tau-b} = \frac{301{,}180 - 14{,}436}{\sqrt{(301{,}180 + 14{,}436 + 1{,}634{,}734)(301{,}180 + 14{,}436 + 176{,}194)}}$$

$$\frac{286{,}744}{\sqrt{(1{,}950{,}350)(491{,}810)}}$$

$$\frac{286{,}744}{\sqrt{959{,}201{,}633{,}500}}$$

$$\frac{286{,}744}{979{,}388.4}$$

$$.293$$

All relationships tested were at the $p < .01$ levels meaning they were significant. However, the p-value only shows that the relationships observed did not occur by chance. The tau-b is used as an inferential statistic to show the strength of those relationships. The following table shows all tau-b values for the relationships tested and from this one can conclude that the relationships are moderate in strength and within an acceptable range.

PAT by GLA- Grade and Subject	***Tau-b***
Gr. 3 Eng., LA	.392
Gr. 6 Eng. LA	.337
Gr. 9 Eng. LA	.293
Gr. 3 Math	.326
Gr. 6 Math	.298
Gr. 9 Math	.303

A primary reason for provincial aggregation of Grade Level of Achievement data is evaluation of education programs such as special education, English as a Second Language, etc. The GLA by PAT analysis demonstrates that GLA data can indeed supplement PAT data with reasonable reliability and validity for the purposes of program evaluation. This observation is particularly relevant for those grades that do not have PAT testing (grades 1, 2, 4, 5, 7, and 8) where GLA can serve as a proxy for PAT data. Additionally, it is useful to be able to supplement PAT data with GLA data in grades 3, 6 and 9 as the added advantage would be broader and richer data to

inform program evaluation related decisions, and to provide data for the approximately 10% of students who for a variety of reasons do not write the PATs.

Further, the fact that the tau-b values show moderate strength lends credibility to the process of collecting GLA. A perfect correlation of 1.0 between GLA and PAT is not an expected nor a desirable condition given the inherent differences underlying the evaluation designs. PAT data are derived from a single paper and pencil test whereas GLA data are based on numerous and more dynamic observations over time, and thus should be a much richer method of assessment, which one could reasonably assume to produce, positively correlated albeit slightly different data than a PAT result.

GLA and Gender

The 2003 analysis of PISA results[8] found that females did much better than males in reading, but males tended to outperform females in mathematics. This pattern of gender differentiation is consistent with the general literature on gender-based test performance differences (Pope, Wentzel and Cammaert, 2002). As another test of the concurrent validity of the GLA data, a gender analysis using Mathematics and English Language Arts means was conducted.

Both Mathematics and English Language Arts data were grouped by male and female, according to grade. Each grade's GLA was totaled, and a mean was calculated. The mean differences between males and females were compared using a T-test for means calculation, and the following tables were produced.

English LA GLA T-Tests

Enrolled Grade	*Gender*	*N*	*Mean GLA*	*Sig.*
1	M	2289	1.00	.394
	F	2203	1.00	
2	M	2759	1.91	.000
	F	2469	1.94	
3	M	2778	2.85	.002
	F	2720	2.88	
4	M	2901	3.80	.001
	F	2693	3.84	
5	M	2869	4.74	.000
	F	2800	4.86	
6	M	2910	5.76	.000
	F	2876	5.85	
7	M	3057	6.80	.000
	F	3076	6.89	
8	M	3094	7.78	.000
	F	2937	7.91	
9	M	2749	8.78	.000
	F	2422	8.88	

[8] PISA 2003 — The 2003 Canadian Report
Measuring Up: Canadian Results of the OECD PISA Study
The performance of Canada's Youth in Mathematics, Reading, Science and Problem Solving
2003 First Findings for Canadians Aged 15

Math GLA T-Tests

Enrolled Grade	*Gender*	*N*	*Mean GLA*	*Sig.*
1	M	2519	1.00	.169
	F	2384	1.01	
2	M	2753	1.96	.418
	F	2468	1.96	
3	M	2739	2.91	.774
	F	2690	2.91	
4	M	2885	3.89	.191
	F	2663	3.90	
5	M	2852	4.84	.000
	F	2768	4.90	
6	M	2876	5.82	.017
	F	2852	5.86	
7	M	3089	6.85	.000
	F	3047	6.91	
8	M	3062	7.85	.000
	F	2914	7.93	
9	M	2734	8.81	.001
	F	2398	8.88	

The above tables show that females outperformed males in English Language Arts by small margins, but the differences were nonetheless statistically significant. The difference between males' and females' mean scores in math were not as pronounced, however they were significant in grades 5 to 9, where females again performed slightly better than males.

The results of the gender analysis of GLA data demonstrate concurrent validity with the 2003 PISA gender based results in language arts. However, the GLA math data, while demonstrating no significant differences between males and females in grades 1-4 do demonstrate that females have significantly higher GLA than do males in grades 5-9.

The GLA results for grades 8 and 9 would be most closely comparable to the PISA data for 15 year olds. The GLA and PISA gender analysis in Mathematics are in opposite directions. This appears to suggest the GLA data lack concurrent validity with the PISA data; however, Pope, Wentzel and Cammaert (2002: 284)) studied the relationship between provincial diploma exam scores and the school awarded mark in all diploma exam subjects and found, "For the school-awarded score results, every course that showed statistically significant gender relationships...had results in the direction of girls outperforming boys." The GLA data reported here demonstrate consistent patterns with the school awarded score data reported in the Pope, Wentzel and Cammaert (2002) study, and may support the hypothesis that there may be "...some sort of differential favoritism in favor of girls in terms of school-awarded scores." These gender relationships are definitely an area worthy of further study both in relationship to GLA data but also in relationship to provincial achievement test data.

Overall Data Observations

The analysis of the Grade Level of Achievement Reporting data was undertaken to assess the validity, reliability and ultimately the utility of the GLA data for judging program impacts. The analysis demonstrated that:

- GLA data, as expected, has a leptokurtic[9] distribution when applied to the general student population, indicating that most students are achieving at grade level. This was also evident in the Spearman correlations between GLA and enrolled grade.
- GLA for sub-groups such as coded students had a greater distribution and wider variance, which increases the utility of the data for judging program impact for these sub-groups.
- The GLA by PAT analysis demonstrated that GLA data can supplement PAT data with reasonable reliability and validity, and with added depth for the purposes of program evaluation. This observation is particularly relevant for those grades that do not have PAT testing where GLA can serve as a proxy for PAT data.
- GLA data provided important data for the approximately 10% of students in grades 3, 6 and 9 who do not write the PATs, thus filling a strategically critical gap in the student achievement database.
- Gender differential analysis of GLA data showed a consistent pattern in relationship to 2002 PISA results for reading, but an inconsistent pattern in mathematics. However, the fact that GLA data demonstrate generally higher scores for girls than boys is consistent with a 2002 study that observed consistently higher high school awarded marks for girls. GLA data will be an important data source for further study of gender-based achievement.
- Most of the data submitted in the first year of the Grade Level of Achievement Reporting Pilot Project were attributed to a jurisdiction that had acquired considerable experience with GLA reporting. This study and the related conclusions will need to be verified when additional jurisdictions' data are available for analysis.

Lessons Learned from the Pilot Study

The stories that have emerged from the jurisdictions involved with the Grade Level of Achievement Reporting pilot (Burger, et. al., 2004) suggest that easy solutions to difficult processes are illusive, and no "one best way" to do classroom assessment exists. In hindsight this is obvious as multiple variables impacted the ability of various schools and jurisdictions to implement a standardized GLA model in the core subjects, ranging from existing assessment knowledge, capacity issues, teacher and administrator "buy-in", methodological issues, educational leadership effectiveness at the school and central office levels, value assumptions, political implications, and of course professional development facility (to name a few). Nonetheless, it was possible to outline rough signposts based on the Grade Level of Achievement Reporting pilot, by which jurisdictions may mark their GLA implementation journey, and offers points for consideration.

First and foremost, it is vital that the implementation be a combination of top down and bottom up efforts. Grade Level of Achievement Reporting proved to be a wonderful exemplar of

[9] As described at http://www.isixsigma.com/dictionary/Leptokurtic_Distribution-268.htm , "A leptokurtic distribution is symmetrical in shape, similar to a normal distribution, but the centre peak is much higher; that is, there is a higher frequency of values near the mean [with resultant reduced variation]."

implementing a specific policy for the sole purpose of solving a specific problem, before those who would benefit most realized a problem existed. The analogy here might be the development of the computer. Originally, computers were massive and complex machines that did little more than mathematical calculations, which were developed for the most part as technical endeavors. It was long after their development that people began to view them as answers to existing data management problems, such as how to track and control increasingly complex air traffic at a growing number of airports. GLA is likewise an incredibly important tool that can and should be included in classroom teachers' and administrators' professional "tool-boxes" to be used in making sense of student achievement information with the view to improving student learning.

Realistically, the task of demonstrating the value of GLA to education professionals is not that difficult. At the root of collecting and reporting GLA is the notion of accountability, and generally when accountability in an educational context is discussed it is often in concert with the notion of standardized testing to inform how student achievement can be positively influenced. Teachers often feel concern about the degree to which a single standardized achievement test can reflect the reality they view in the classrooms, consequently teachers may see a benefit in having their classroom assessment become part of a more comprehensive public discourse on what is useful school and jurisdiction evaluation. The Grade Level of Achievement Reporting pilot seemed to avail this opportunity, and many of the jurisdiction representatives felt the pilot was valuable in this regard, and in the potential to enrich the dialogue between school-based and central office based staff on student achievement matters.

Having said this however, it was also apparent in the Grade Level of Achievement Reporting pilot that there was a degree of apprehension in regards to reporting GLA, presumably because teachers in the pilot jurisdictions felt somewhat vulnerable concerning being accountable for their assessments. While understandable within the context of traditional education contexts, emerging models of educational leadership and effective schools point to more open systems where students and parents are much more engaged in assessment for learning approaches (Stiggins, 2001) and the public is much better informed about what works in schools (Reeves, 2004).

In the participating jurisdictions there seemed to be individuals willing to act as policy "entrepreneurs" who were prepared to become "advocates", and convince others in their professional communities that there was merit in the GLA Reporting process. The policy entrepreneur's role may simply be a matter of reassuring teachers that the type of accountability Grade Level of Achievement Reporting seeks to serve is not teacher accountability, but learner-centered accountability in order to serve the students' needs. Further, teachers may see this as an opportunity to boost their professionalism and confidence, as they are the ones directly in control of the accountability indicators in such a model (Reeves: 2004).

In short, the main themes that emerged from the Grade Level of Achievement Reporting pilot were:

1. Implementing a process for collecting GLA works best when it is related to existing assessment work- The reasons for this seem to be:
    a. Teachers do not need more work to do, and if it seems like that is what it will be, the chance of getting "buy in" will be minimal. Rather, they need to view it as an important aspect of work they are already undertaking.

b. If attached to existing assessment work, there seems to be a pre-existing value for the role of assessment as a formative tool that improves teaching and learning outcomes. Teachers will be more likely to see the worth of GLA if they already value the role of assessment as a formative process that ultimately drives improvement in the summative side of classroom assessment.
c. If it is attached to existing work, it does not seem imposed. There is an intrinsic value to letting people discover and internalize the value of good, well-balanced assessment tools, and the benefits they confer.
d. Working backwards from clear learning objectives assessed with well designed performance assessment to define GLA is a powerful approach to improved pedagogy.
e. Grade level of achievement data driven by formative, assessment for learning methods is a reasonable source of information, with acceptable concurrent and predictive validity, that is useful for informing summative judgments of program impacts at the classroom, school, jurisdiction and provincial levels and hence has a maximum degree of usefulness.
f. Classroom generated grade level of achievement data adds value and contributes to our knowledge of the existing data collected for schools and jurisdictions from provincial achievement tests and provides a more comprehensive picture of student achievement.
g. The process of generating grade level of achievement data for reporting has potential for having a positive impact on teacher professional growth and pedagogy.

2. The role of policy entrepreneurs at multiple levels is vital. Simply agreeing to do it does not mean it will be successful. First, it takes people who believe in a project like this, and are willing to act as advocates and go into the schools or central offices to advocate for GLA Reporting. Having individuals in jurisdictions that are willing to work with their professional learning communities and see the opportunities rather than the risks associated with reporting GLA, helps mitigate the anxieties some may have regarding accountability. Further, even if people agree to initially become part of the process; it likely will not be a success if assessment expertise and some positive reinforcements are not available to them at the school level. In order to be a success, champions of assessment for learning must network and communicate actively from start to finish.

3. Grassroots- It seems to work best as a ground-up model with maximum opportunities for teacher participation and ownership of the classroom assessment process, facilitated and aided by a committed administration.

4. Nobody works in isolation, nobody expects to be perfect- The pilot jurisdictions seemed to be of the opinion that they would benefit from the Grade Level of Achievement Reporting project, but also viewed it as a foundational step towards GLA data generation in support of broad-based improvement and reporting strategies. In other words, many viewed it as a project of discovery, or an initial stage of a process that will undoubtedly be refined as it evolves. It is vital that jurisdictions do not feel they can simply implement the policy and immediately reap the benefits. The pilot jurisdictions largely felt it is a process that needs cultivation and encouragement over time in order to see the full-range of positive impacts. Furthermore jurisdiction staff were willing to show a

degree of vulnerability in agreeing to participate in the pilot and in asking other jurisdiction staff for help.

5. Lastly, individual teachers in the classrooms benefit from experiencing the benefits of collegial networking and discussions of how to best judge a student's grade level of achievement. Teachers and administrators also need to see that the value of GLA data applied within a context of professional learning communities characterized by critical reflection far outweighs any possible risks.

Conclusion

Alberta Education in the spring of 2005 announced that Grade Level of Achievement Reporting would be phased in province-wide over a three year period culminating with GLA data being reported for all grade 1-9 students in the four core subjects by June 2008. Planning is well underway within Alberta Education to optimize the usefulness of the GLA data for school, jurisdiction and provincial level analysis and decision-making. Planning is also well underway to provide key supports to teachers and administrators in reporting GLA data.

From one perspective, making judgements about GLA is seen as fundamental to good teaching and hence is likely well grounded in most schools given that knowing a student's instructional level is a fundamental prerequisite to effective teaching. Alternatively, the view has also been advanced that additional teacher supports will need to be developed and made available. To that end Alberta Education has contracted the Alberta Assessment Consortium to develop a Guide to GLA Reporting that will reference helpful methodological concepts and a wide range of assessment supports available or under development. Consultations are also well underway with the Alberta faculties of education regarding what optimal approaches to enhancing pre-service teaching of classroom assessment knowledge and skills entails. Lastly, Alberta Education is considering how in-service can best support the GLA Reporting initiative.

The vision of accountability and program evaluation in the future represented in this paper is one of a rich environment where assessment of learning and assessment for learning are complementary approaches in a comprehensive assessment model. Rich data will assist professional learning communities to be connected across organizational boundaries, and educational leadership will be reflected in a complex and connected network of professionals informed with timely and comprehensive information and data flows that result in the program and programming decisions that improve the quality of education available to students. Grade Level of Achievement Reporting is one small step in this direction.

Bibliography

Alberta Learning, (2002). *Standards for Special Education.* Edmonton, AB. Author.

Alberta Learning (2001). *Entry Age, Age Within Cohort, and Achievement.* Edmonton, AB. Author.

Alberta Learning (2001). *MIRS Next Generation: Design Principles for a Learner Results Database for Improved Program Evaluation.* Edmonton, AB. Author.

Archer, K., Gibbins, R., and Youngman, L. (1998). Explorations: A Navigator's Guide to Quantitative Research in Canadian Political Science

Bloom, B. *(1980).* The New Direction in Education Research: Alterable Variables. *Phi Delta Kappan. V.61 N.6.*

Burger, J. et. al. (2004). "Accountability for Learning: The Beyond MIRS Pilot Project." Paper presented at the Alberta Assessment Consortium Conference. Fantasyland Hotel, Edmonton, AB. October 23, 2004.

Burger, J., and Krueger, M. (2003). "A Balanced Approach to High-Stakes Achievement Testing: An analysis of the literature with policy implications" *International Electronic Journal for Leadership in Learning.* University of Calgary.

Earl, L., and Katz, S. (2002). "Leading Schools in a Data-Rich World" in Leithwood, K., and Hallinger, P., (Eds.), *Second International Handbook of Educational Leadership and Administration, VI.* Kluwer Academic Publishers. Dordrecht, Netherlands.

Louis, K. S., Febey, K, and Schroeder, R.(2005). "State-Mandated Accountability in High Schools: Teachers' Interpretations of a New Era." *Educational Evaluation and Policy Analysis.* Vol. 27(4), 177-204.

OECD PISA (2003). *Measuring Up: Canadian Results of the OECD PISA Study The performance of Canada's Youth in Mathematics, Reading, Science and Problem Solving 2003 First Findings for Canadians Aged 15.*

Pope, G.A., Wentzel, C., & Cammaert, R. (2003). Relationships between gender and Alberta diploma scores. *Alberta Journal of Educational Research,* 48(4), 275-286

Reeves, D. B. (2004) *Accountability For Learning: How Teachers and School Leaders Can Take Charge.* Assoc. for Supervision and Curriculum Development, Alexandria, VA.

Stecher, B., and Kirby, S. N. (2004). Organizational Improvement and Accountability: Lessons for Education From Other Sectors

Stiggins, R. J. (2001) *Student Involved Classroom Assessment,* 3rd. Edition. Merrill Prentice Hall. Columbus, Ohio.